



Data Mining and Machine Learning for Political Events

An Application to Produce Indicators for Presidential Systems

Bastián González-Bustamante

University of Oxford

✉ bastian.gonzalezbustamante@politics.ox.ac.uk

Presentation at the ECPR General Conference, Charles University
Prague, Czech Republic, September 4-8, 2023

Table of Contents

1. Introduction
2. Creation of the Dataset
3. Machine Learning Classifiers
4. Assessing Dataset Validity
5. Potential Applications
6. Take Aways



Introduction

Data Mining and Machine Learning

Employing a strategy based on **Natural Language Processing** (NLP) and a **text-as-data approach**, this paper presents an entirely novel dataset constructed by applying data mining and machine learning techniques.

The dataset contains detailed information on cabinet turnover in **12 presidential cabinets in Latin America from the mid-1970s to the early 2020s**. The indicators on resignation calls and reallocations of cabinet members are entirely novel.

I focus on the creation of the dataset using Optical Character Recognition (OCR) on press report archives together with machine learning models. The models permitted the training of ensemble **semi-supervised classifiers over a period of almost 50 years**.

Creation of the Dataset

Creation of the Dataset

I combined the **Tesseract OCR** (Ooms, 2021) with different **semi-supervised machine learning models** to distinguish between resignation calls and other mentions of cabinet members in the press.

This was applied to **Latin American Weekly Report (LAWR) archives**.



Detail of the Batches of LAWR Archives

Batch	Archives	Years	Format	Volume
Batch-01	LAWR-1975 to LAWR-1979 Including 1997	5	 CSV UTF-8	9.56 MB
Batch-02	LAWR-1980 to LAWR-1998 Excluding 1997 Including the first half of 2003	18.5	 PNG images	38.5 GB
Batch-03	LAWR-1999 to LAWR-2021 Excluding the first half of 2003	22.5	 PDF files	341 MB

Binarisation Techniques for Enhancing OCR

POSTSCRIPT	
POLITICS	VENEZUELA CABINET
President: Hugo Chávez	Finance: Pedro Rojas Interior: Víctor Antón Agriculture: Roberto López Health: Endre Pötter Education: Carlos Gómez Labour: Francisco Márquez Development: Julio Carrasco Transport and communications: Carmen González Urban development: César Gómez Justice: José Pérez Lorenzo Foreign Affairs: Humberto Bell Agriculture: Roberto López Transport and communications: Endre Pötter Development: Julio Carrasco Labour: Francisco Márquez Central Bank: Silvia de Gutiérrez Attorney general: Guillermo Vargas Bustamante
Deputy President: U N. de Alvarado	Deputy: Jaime Rodríguez Justice: José Pérez Lorenzo Foreign Affairs: Humberto Bell Agriculture: Roberto López Transport and communications: Endre Pötter Development: Julio Carrasco Labour: Francisco Márquez Central Bank: Silvia de Gutiérrez Attorney general: Guillermo Vargas Bustamante
Ministers: 11	Deputy Minister: 1
Total: 12	Total: 12

PERUVIAN VIOLENCE	
Wk	Mo
1980	125
1981	125
1982	145
1983	1,450
1984	1,521
1985	1,545
1986	1,521
1987	1,545
1988	1,521
1989	1,545
1990	1,521
1991	1,545
1992	1,521
1993	1,545
1994	1,521
1995	1,545
1996	1,521
1997	1,545
1998	1,521
1999	1,545
2000	1,521
2001	1,545
2002	1,521
2003	1,545
2004	1,521
2005	1,545
2006	1,521
2007	1,545
2008	1,521
2009	1,545
2010	1,521
2011	1,545
2012	1,521
2013	1,545
2014	1,521
2015	1,545
2016	1,521
2017	1,545
2018	1,521
2019	1,545
2020	1,521
2021	1,545
2022	1,521
2023	1,545
2024	1,521
2025	1,545
2026	1,521
2027	1,545
2028	1,521
2029	1,545
2030	1,521
2031	1,545
2032	1,521
2033	1,545
2034	1,521
2035	1,545
2036	1,521
2037	1,545
2038	1,521
2039	1,545
2040	1,521
2041	1,545
2042	1,521
2043	1,545
2044	1,521
2045	1,545
2046	1,521
2047	1,545
2048	1,521
2049	1,545
2050	1,521
2051	1,545
2052	1,521
2053	1,545
2054	1,521
2055	1,545
2056	1,521
2057	1,545
2058	1,521
2059	1,545
2060	1,521
2061	1,545
2062	1,521
2063	1,545
2064	1,521
2065	1,545
2066	1,521
2067	1,545
2068	1,521
2069	1,545
2070	1,521
2071	1,545
2072	1,521
2073	1,545
2074	1,521
2075	1,545
2076	1,521
2077	1,545
2078	1,521
2079	1,545
2080	1,521
2081	1,545
2082	1,521
2083	1,545
2084	1,521
2085	1,545
2086	1,521
2087	1,545
2088	1,521
2089	1,545
2090	1,521
2091	1,545
2092	1,521
2093	1,545
2094	1,521
2095	1,545
2096	1,521
2097	1,545
2098	1,521
2099	1,545
2000	1,521
2001	1,545
2002	1,521
2003	1,545
2004	1,521
2005	1,545
2006	1,521
2007	1,545
2008	1,521
2009	1,545
2010	1,521
2011	1,545
2012	1,521
2013	1,545
2014	1,521
2015	1,545
2016	1,521
2017	1,545
2018	1,521
2019	1,545
2020	1,521
2021	1,545
2022	1,521
2023	1,545
2024	1,521
2025	1,545
2026	1,521
2027	1,545
2028	1,521
2029	1,545
2030	1,521
2031	1,545
2032	1,521
2033	1,545
2034	1,521
2035	1,545
2036	1,521
2037	1,545
2038	1,521
2039	1,545
2040	1,521
2041	1,545
2042	1,521
2043	1,545
2044	1,521
2045	1,545
2046	1,521
2047	1,545
2048	1,521
2049	1,545
2050	1,521
2051	1,545
2052	1,521
2053	1,545
2054	1,521
2055	1,545
2056	1,521
2057	1,545
2058	1,521
2059	1,545
2060	1,521
2061	1,545
2062	1,521
2063	1,545
2064	1,521
2065	1,545
2066	1,521
2067	1,545
2068	1,521
2069	1,545
2070	1,521
2071	1,545
2072	1,521
2073	1,545
2074	1,521
2075	1,545
2076	1,521
2077	1,545
2078	1,521
2079	1,545
2080	1,521
2081	1,545
2082	1,521
2083	1,545
2084	1,521
2085	1,545
2086	1,521
2087	1,545
2088	1,521
2089	1,545
2090	1,521
2091	1,545
2092	1,521
2093	1,545
2094	1,521
2095	1,545
2096	1,521
2097	1,545
2098	1,521
2099	1,545
2000	1,521
2001	1,545
2002	1,521
2003	1,545
2004	1,521
2005	1,545
2006	1,521
2007	1,545
2008	1,521
2009	1,545
2010	1,521
2011	1,545
2012	1,521
2013	1,545
2014	1,521
2015	1,545
2016	1,521
2017	1,545
2018	1,521
2019	1,545
2020	1,521
2021	1,545
2022	1,521
2023	1,545
2024	1,521
2025	1,545
2026	1,521
2027	1,545
2028	1,521
2029	1,545
2030	1,521
2031	1,545
2032	1,521
2033	1,545
2034	1,521
2035	1,545
2036	1,521
2037	1,545
2038	1,521
2039	1,545
2040	1,521
2041	1,545
2042	1,521
2043	1,545
2044	1,521
2045	1,545
2046	1,521
2047	1,545
2048	1,521
2049	1,545
2050	1,521
2051	1,545
2052	1,521
2053	1,545
2054	1,521
2055	1,545
2056	1,521
2057	1,545
2058	1,521
2059	1,545
2060	1,521
2061	1,545
2062	1,521
2063	1,545
2064	1,521
2065	1,545
2066	1,521
2067	1,545
2068	1,521
2069	1,545
2070	1,521
2071	1,545
2072	1,521
2073	1,545
2074	1,521
2075	1,545
2076	1,521
2077	1,545
2078	1,521
2079	1,545
2080	1,521
2081	1,545
2082	1,521
2083	1,545
2084	1,521
2085	1,545
2086	1,521
2087	1,545
2088	1,521
2089	1,545
2090	1,521
2091	1,545
2092	1,521
2093	1,545
2094	1,521
2095	1,545
2096	1,521
2097	1,545
2098	1,521
2099	1,545
2000	1,521
2001	1,545
2002	1,521
2003	1,545
2004	1,521
2005	1,545
2006	1,521
2007	1,545
2008	1,521
2009	1,545
2010	1,521
2011	1,545
2012	1,521
2013	1,545
2014	1,521
2015	1,545
2016	1,521
2017	1,545
2018	1,521
2019	1,545
2020	1,521
2021	1,545
2022	1,521
2023	1,545
2024	1,521
2025	1,545
2026	1,521
2027	1,545
2028	1,521
2029	1,545
2030	1,521
2031	1,545
2032	1,521
2033	1,545
2034	1,521
2035	1,545
2036	1,521
2037	1,545
2038	1,521
2039	1,545
2040	1,521
2041	1,545
2042	1,521
2043	1,545
2044	1,521
2045	1,545
2046	1,521
2047	1,545
2048	1,521
2049	1,545
2050	1,521
2051	1,545
2052	1,521
2053	1,545
2054	1,521
2055	1,545
2056	1,521
2057	1,545
2058	1,521
2059	1,545
2060	1,521
2061	1,545
2062	1,521
2063	1,545
2064	1,521
2065	1,545
2066	1,521
2067	1,545
2068	1,521
2069	1,545
2070	1,521
2071	1,545
2072	1,521
2073	1,545
2074	1,521
2075	1,545
2076	1,521
2077	1,545
2078	1,521
2079	1,545
2080	1,521
2081	1,545
2082	1,521
2083	1,545
2084	1,521
2085	1,545
2086	1,521
2087	1,545
2088	1,521
2089	1,545
2090	1,521
2091	1,545
2092	1,521
2093	1,545
2094	1,521
2095	1,545
2096	1,521
2097	1,545
2098	1,521
2099	1,545
2000	1,521
2001	1,545
2002	1,521
2003	1,545
2004	1,521
2005	1,545
2006	1,521
2007	1,545
2008	1,521
2009	1,545
2010	1,521
2011	1,545
2012	1,521
2013	1,545
2014	1,521
2015	1,545
2016	1,521
2017	1,545
2018	1,521
2019	1,545
2020	1,521
2021	1,545
2022	1,521
2023	1,545
2024	1,521
2025	1,545
2026	1,521
2027	1,545
2028	1,521
2029	1,545
2030	1,521
2031	1,545
2032	1,521
2033	1,545
2034	1,521
2035	1,545
2036	1,521
2037	1,545
2038	1,521
2039	1,545
2040	1,521
2041	1,545
2042	1,521
2043	1,545
2044	1,521
2045	1,545
2046	1,521
2047	1,545
2048	1,521
2049	1,545
2050	1,521
2051	1,545
2052	1,521
2053	1,545
2054	1,521
2055	1,545
2056	1,521
2057	1,545
2058	1,521
2059	1,545
2060	1,521
2061	1,545
2062	1,521
2063	1,545
2064	1,521
2065	1,545
2066	1,521
2067	1,545
2068	1,521
2069	1,545
2070	1,521
2071	1,545
2072	1,521
2073	1,545
2074	1,521

Tesseract's Layout Analysis Module

Latin American weekly report

How Domingo Cavallo rose to become Menem's virtual prime minister

Argentina is still holding debates and lay the blame for the military coup on former minister of Interior Ernesto González and his replacement by Domingo Cavallo (MAY-91). The simple explanation is that the coup was carried out by a group of older state, as it is dismissed as sufficient reason, and much has been written about all sorts of political plots.

One thing seems certain: that General Menéndez, like his predecessor at Ministry of Interior, according to sources with access to him, that he knew President Carlos Menem had been plotting to oust him, a well-accepted version among those that Menéndez had always intended to become President of Argentina, to succeed to the period of inactivation, which he expected to be brief. That plan, according to sources, was to choose in his own favor, the man who would succeed him.

PERU: Spread of Chagas epidemic. (2)

VENEZUELA: "Holiday" education impact on wife. (2)

COLOMBIA: Prospects of new peace deal. (2)

VENDEZUELA: Lower prices force budget rethink. (2)

BOLIVIA: A Soviet partner for the oil industry. (2)

PARAGUAY: Reactions to democracy. (4)

PERU: Watch out for the "holiday" education. (2)

BRAZIL: Five friends for Team Brazil. (2) (3)

CHILE: Proposals to extend Aymara's mandate. (2)

HAITI: Aristide faces first major challenges. (2)

EL SALVADOR: Getback as FMLN threatens secession. (2)

COSTA RICA: Preparing a law just in case. (1)

PANAMA: Time to recall the old guard. (1)

The image shows the front cover of a publication titled "Latin American weekly report". The title is written in a large, bold, sans-serif font. Above the title, there is a black rectangular box containing the word "Latin" in white. To the right of the title, there is a large graphic of the United Nations emblem, which consists of a globe surrounded by a ring of 18 horizontal bars of equal width, alternating in light blue and light yellow. Below the title, the date "21 January 1987" is printed. At the bottom of the cover, there is some smaller text and a small logo.

How Domingo Cavallo rose to become Menem's virtual prime minister

Argentines are still hotly debating what lay behind the sudden resignation of economy minister Aníbal Fernández González and his replacement by Domingo Cavallo (MNR-06). The simple explanation—rapid, uncontrolled upside of the dollar rate—is dismissed as a spurious reason, and much has been published about all sorts of political plots.

One thing seems certain: González had several times warned Menem of his resignation at Menem—and, according to sources with access to him, that he knew President Carlos Menem would one day accept. Indeed, a widely accepted version is that Menem had always intended to appoint Cavallo as his economy minister, to preside over the period of 'inactivation' which he expected would follow a first phase of 'stabilisation'. In this scenario, all González did was to choose his own timeline.

PERU: Wildfires spread of cholera epidemic, [B]	VENEZUELA: A. Lower prices force rural families to migrate, [S]	BRAZIL: Five friends for Piano Cellor, [B]	EL SALVADOR: Setback as FMLN leadership secret, [B/10]
EQUADOR: "Holiday" recession impact on wine, [S]	BOLIVIA: A. Soviet partner for oil exploration, [B]	CHILE: Proposes to extend Assange's mandate, [S]	COSTA RICA: Preparing a law just in case, [H]
COLOMBIA: Prospects of peace deal with FARC, [B]	PANAMA: A. Tightening limits on democracy, [B]	HAITI: Assange faces first major challenges, [S]	PANAMA: Time to recall
URUGUAY: New president, [B]	REGRAN: Watch out for the new president, [B]	PAKISTAN: A. US drone strike kills 10, [B]	PAKISTAN: US drone strike kills 10, [B]

weekly report

How Domingo Cavallo rose to become Menem's virtual prime minister

Argentines are still hotly debating what lay behind the sudden resignation of economy minister Andrés Emilio González and his replacement by Domingo Cavallo (WPL 08). The simple explanation — a rapid, uncontrolled upsurge of the dollar rate — is dismissed as sufficient reason, and much has been published about all sorts of political plots.

One thing seems certain: the

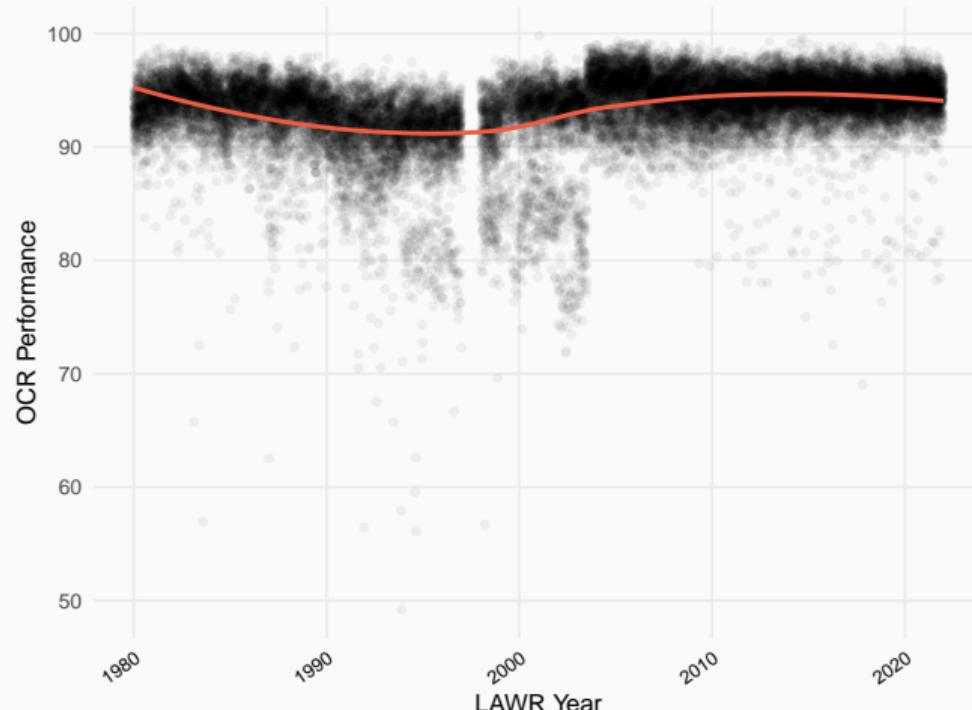
González had several times was his resignation at Menem—and, according to sources, with access him, that he knew President Carlos Menem would one day accept. Instead, a widely accepted version that Menem had always intended appoint Cavallo as his economy minister, to preside over the period "activation" which he expects would follow a first phase of "staturation". In this scenario, all González did was to choose his own timing.

- PERU: Whistleblower spread of cholera epidemic, [2]
- ECUADOR: "Holiday" illnesses impact on strike, [2]
- COLOMBIA: Prospects of life above board with IMF, [2]

A: Lower prices at re-think.	BRAZIL: Free trade deal for Paraná Callor, (3/13)	EL SALVADOR: Setback FMNL 'Timeline secretary', (18/13)
Soviet partner statement, (2) Planning limita- tions, (4)	CHILE: Proposal to extend Aylwin's mandate, (9)	COSTA RICA: Preparing a law just in case, (11)
Trade out for the rest of us, (2)	HAITI: Aristide faces final major challenges,	PANAMA: Time to recall Panama's 'missives', (13)
U.S. policy, (2)	PALESTINE: The newly elected leadership, (1)	

B González-Bustamante

Performance and Accuracy of the Tesseract OCR Algorithm



Accuracy is based on the proportion of text that was correctly identified.

This calculation was performed after extracting stop words and carried out with dictionaries trained with **Long Short-Term Memory (LSTM) models** used by Tesseract.

Machine Learning Classifiers

Semi-Supervised Approach

There are two significant **drawbacks** to training **supervised algorithms**:

1. This would have required a major amount of labelled text data (**time and resources**)
2. The accuracy of the algorithms is expected to decrease out-of-the-sample and over time
(batches of archives cover almost 50 years)

Therefore, I preferred a **semi-supervised approach** that could maintain a **high accuracy level over time with a small amount of coded data** for training (Lee et al., 2021).

1. I drew a subsample of 1,000 LAWR press releases
2. A team of human coders classified all these observations (**resignation calls**)
3. Labelling was conducted to generate ground-truth labels to cross-validate
4. Semi-supervised models using knowledge-based seed words

Semi-Supervised Approach

Pre-processing

80% of the subsample for training

20% for validation

100 tokens per news item *

SMART stop words **

Customised list ***

Stemming algorithm

* until 500 for robustness checks

** Snowball for robustness checks

*** based on LAWR numbering

Models

Naïve Bayes (NB)

Support Vector Machine (SVM) *

Radial Basis Function (RBF) SVM **

Random Forest with 100 trees (RF100)

Random Forest with 500 trees (RF500)

Extreme Gradient Boosting (XGBoost)

* Kernel-based Linear

** Kernel-based Gaussian

Labelling Process for Benchmarking Supervised Models

Coder	Labels	$M(T)$ Label	\sum Time	Consensus
1	966	39s	10h 25m	94.05
2	718	1m 15s	15h 0m	94.84
3	704	38s	7h 29m	94.82
4	318	1m 4s	5h 38m	95.47
5	263	17s	1h 14m	99.26
6	101	1m 17s	2h 9m	88.50

Note. Each observation was hand-coded by three different coders. 2.3% of observations were revised and relabelled during an iteration step. Seed words coincide with features extracted from the benchmarking supervised models trained based on this labelling process.

Goodness-of-Prediction of the Semi-Supervised Models

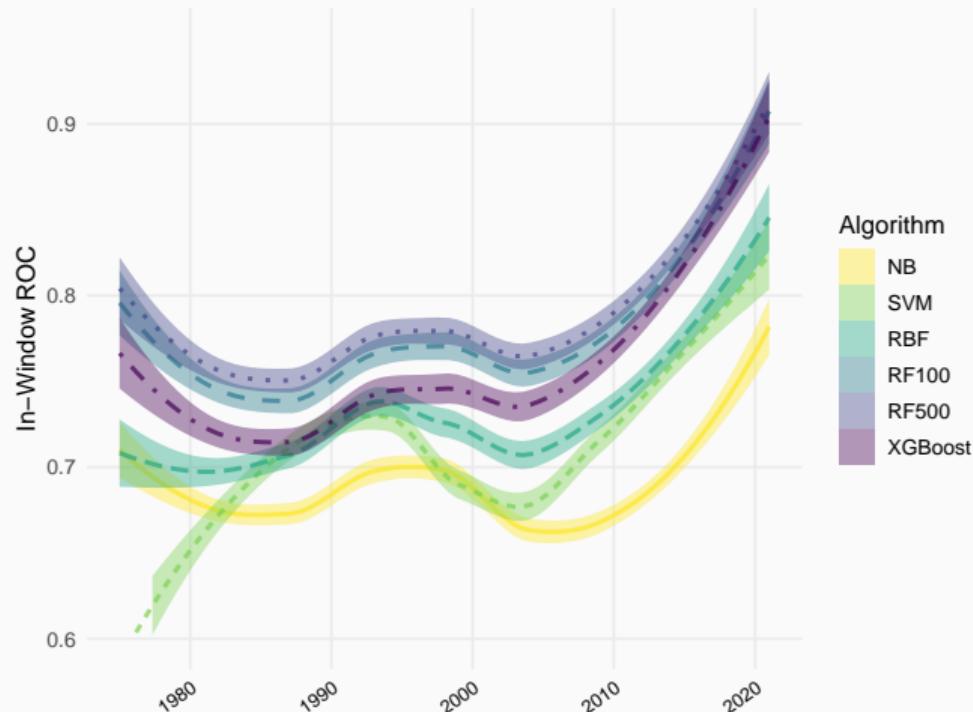
Model	Knowledge-Based Seed Words				
	Accuracy	Precision	Recall	F1-score	ROC
Naïve Bayes	0.713	0.734	0.948	0.828	0.638
Kernel Linear SVM	0.698	0.741	0.900	0.812	0.614
Kernel Gaussian RBF SVM	0.726	0.728	0.995	0.840	0.647
Random Forest (100 trees)	0.743	0.742	0.991	0.849	0.671
Random Forest (500 trees)	0.743	0.740	0.997	0.849	0.682
XGBoost	0.726	0.762	0.905	0.827	0.660

Note. The evaluation was performed with resampling methods using the testing set and 10-fold cross-validation.

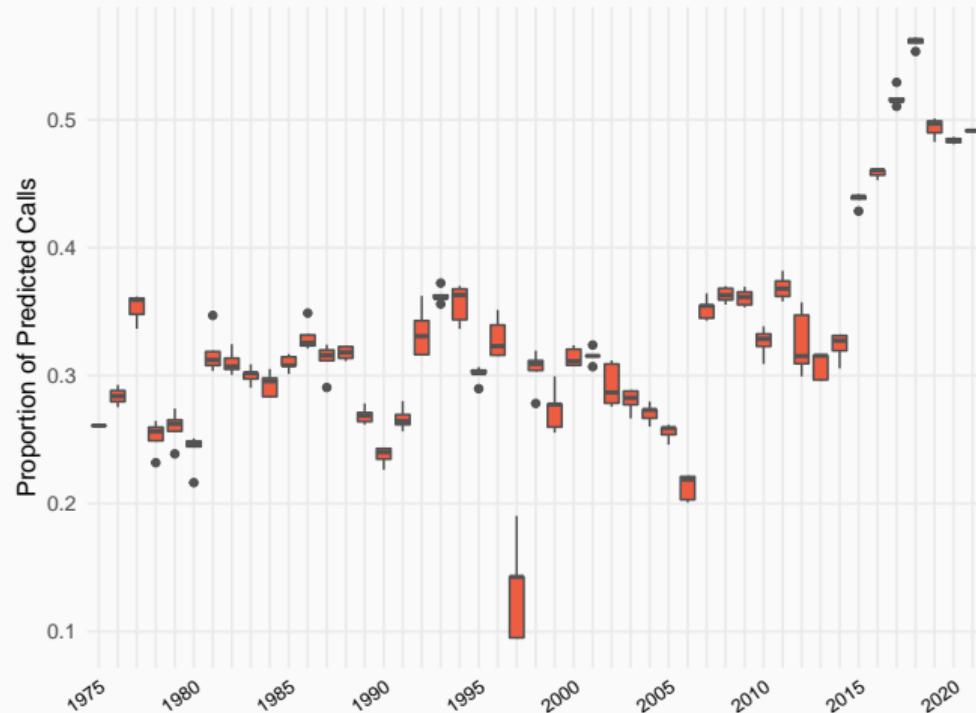
ROC Semi-Supervised Models to Classify Ministerial Resignation Calls

Following a strategy similar to Greene et al. (2019), I used a **five-year fixed rolling window**.

ROC indicators were 10-fold cross-validated in the five-year rolling window testing set between 1975 and 2021.



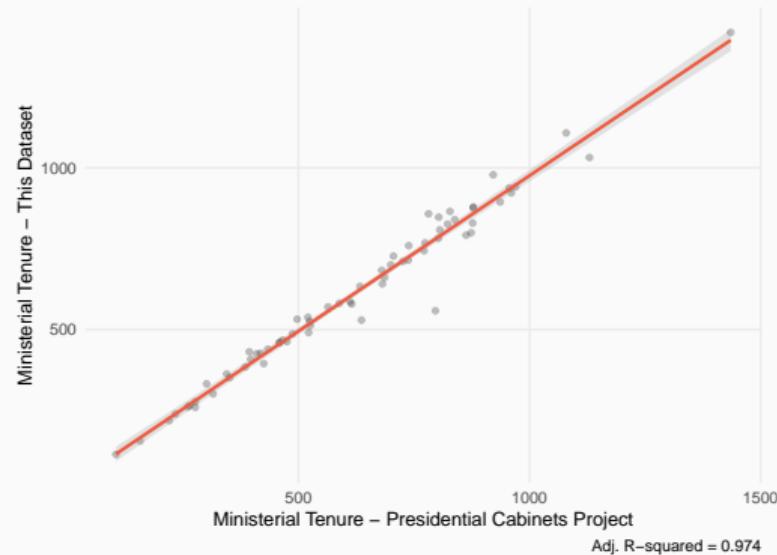
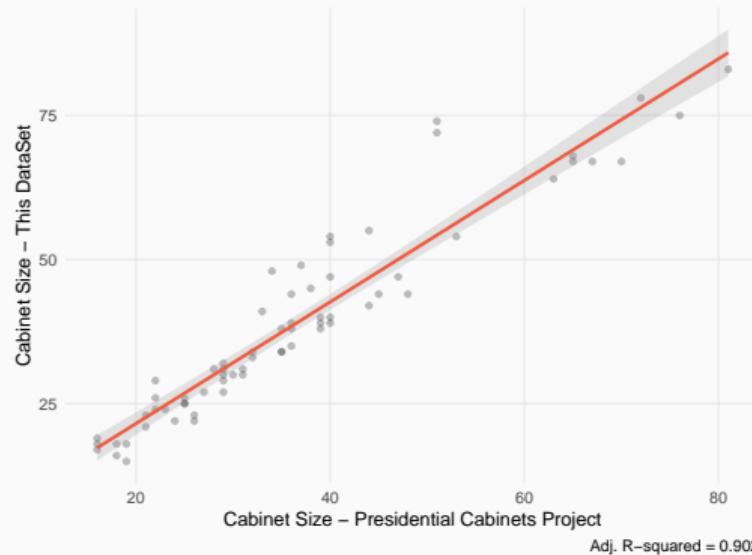
Predicted Ministerial Resignation Calls over Cabinet Press Mentions



Proportion of predicted resignation calls over mentions of cabinet members in press reports, were carried out with an **ensemble semi-supervised RF500 algorithm**.

Assessing Dataset Validity

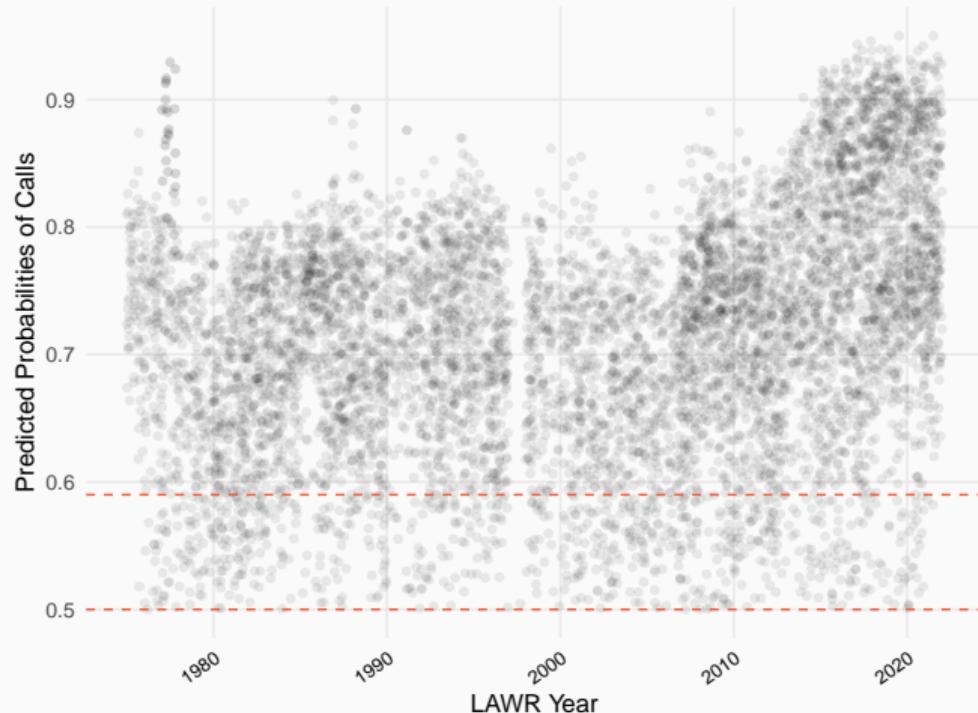
Plausibility Checks against Similar Datasets



Note. Concurrent presidential terms were used for comparisons with the Presidential Cabinet Project (see [Camerlo and Martínez-Gallardo, 2020](#)).

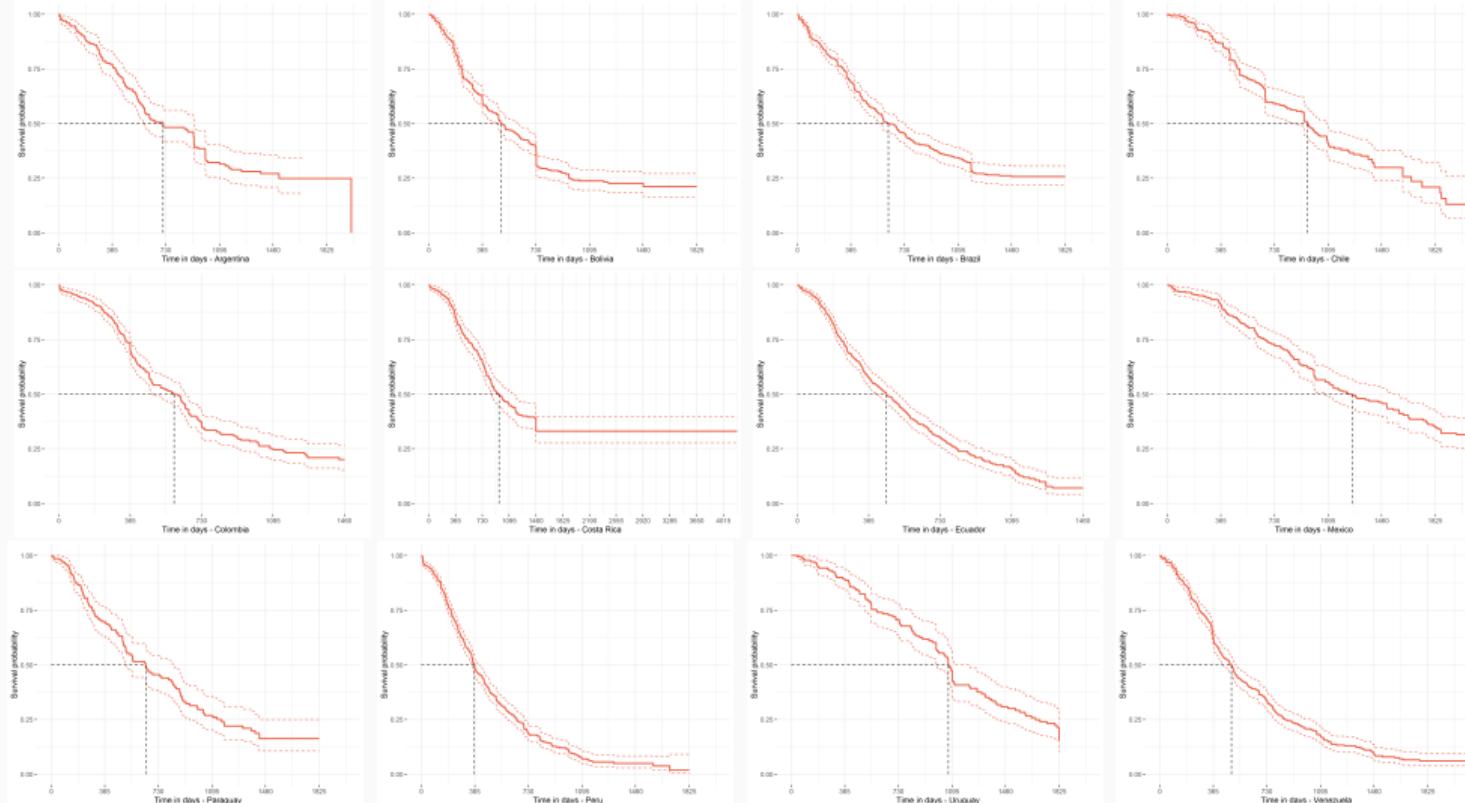
Augmented Artificial Intelligence

I performed an augmented artificial intelligence process based on a **human review** of the predicted observations with low confidence values to improve the data's quality.

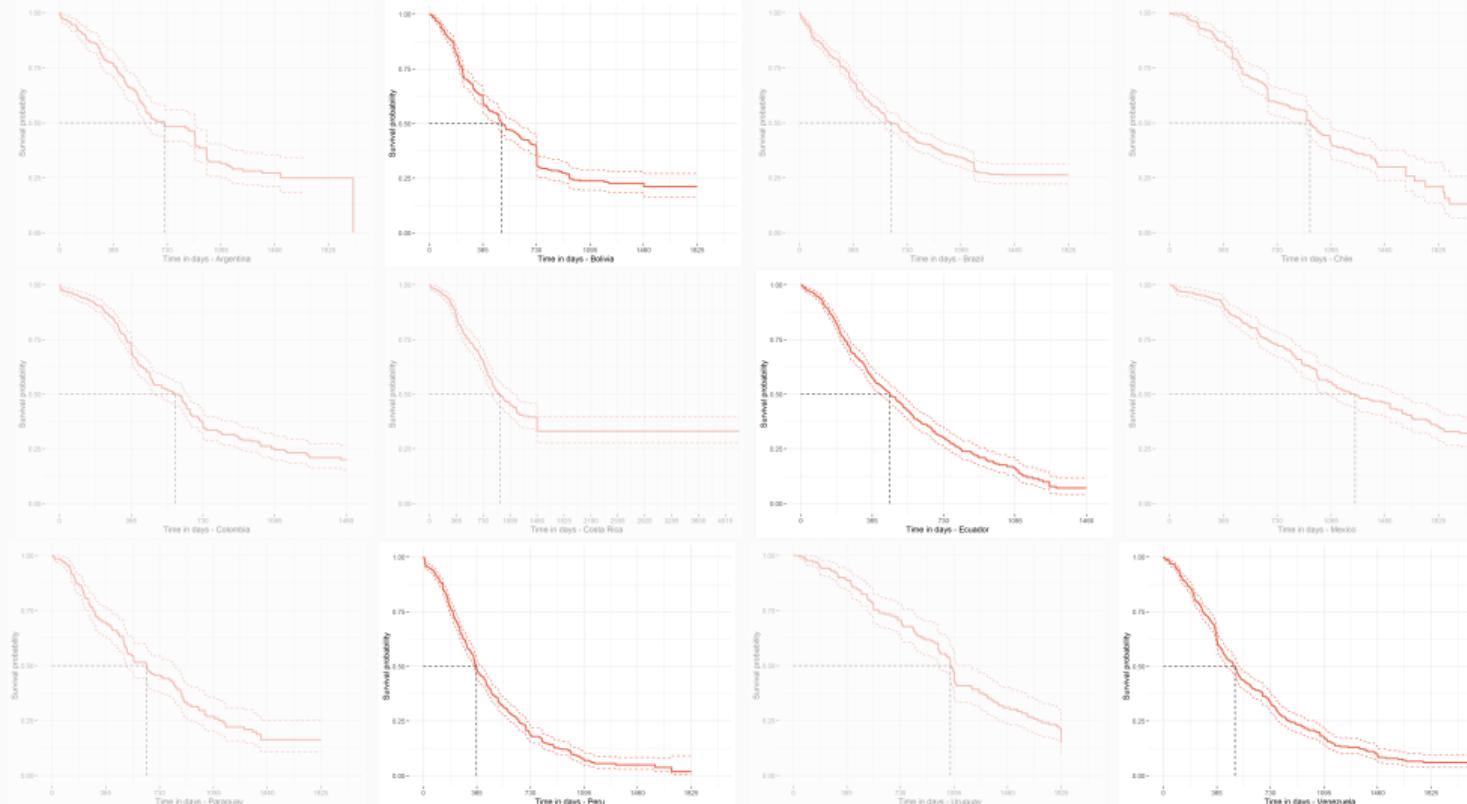


Potential Applications

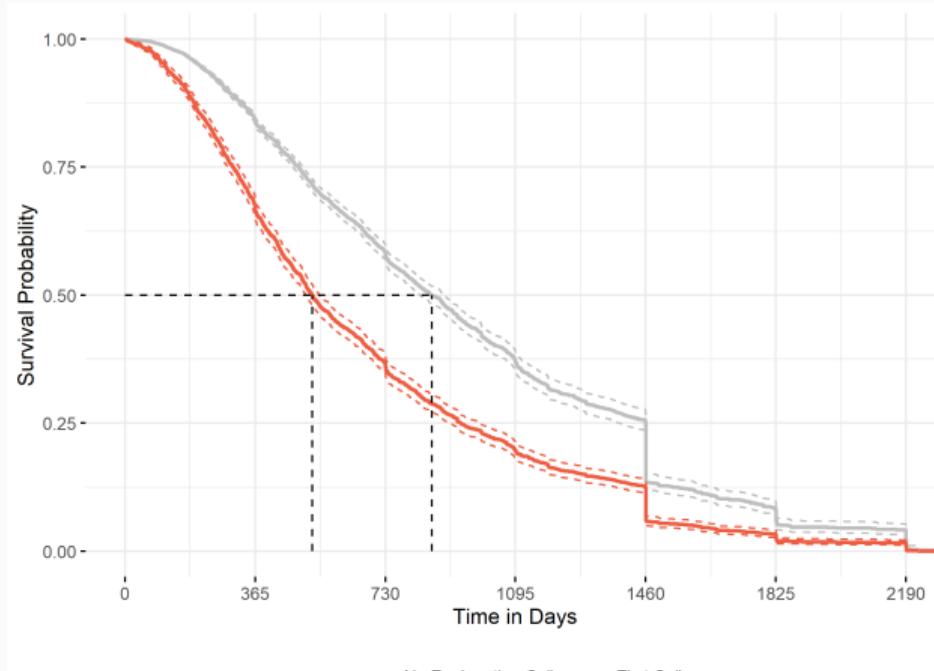
Kaplan-Meier Survival Estimations in 12 Presidential Cabinets



Kaplan-Meier Survival Estimations in 12 Presidential Cabinets



Kaplan-Meier Survival Estimation

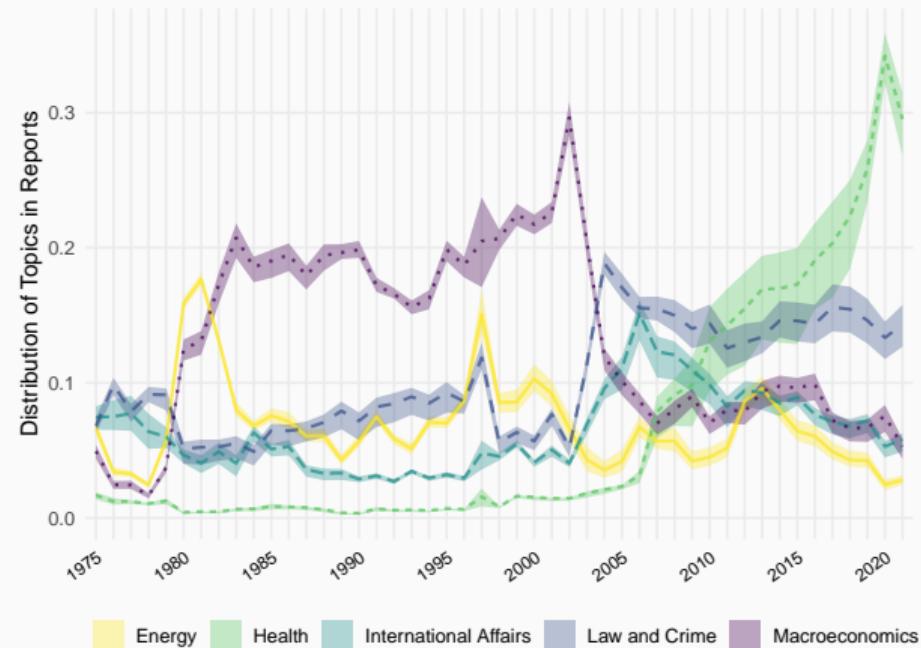


$F(t) = 1 - S(t) = \Pr(T \leq t)$ can be used to obtain the incidence of ministerial exits during T by considering $S(t)$ as a survival function, if raw exit is considered an event in the **absence of competing risks**.

Evolution of Major Agenda Topics in Latin America

I used **Keyword Assisted Topic Modelling** (**key-ATM**; see Eshima et al., 2022) and a dictionary based on Comparative Agendas Project (2022).

It is a semi-supervised Bayesian application that allows me to obtain ϑ (distribution of topics in the documents and Cls).



Take Aways

Take Aways

- To date, this is one of the **most extensive datasets available on presidential cabinets**, going beyond most existing data that tend to focus only on reshuffles or individual ministerial terminations.
- With semi-supervised models, **a high level of accuracy** can be obtained from a **small amount of training data** (Lee et al., 2021; Watanabe, 2018; Watanabe and Zhou, 2022).
- Additional manual labelling work would be necessary to **identify the source of the questioning** (i.e., opposition leaders, internal disagreements, etc.) This opens new avenues for research on specific dynamics depending on the type of scandal.
- Considering the impressive development and leap forward in AI during recent years, **new classifiers could be trained** using generate transformers and LLMs.

References

References I

- Camerlo, M. and Martínez-Gallardo, C. (2020). The Presidential Cabinets Project. Research project, Universidade de Lisboa and University of North Carolina at Chapel Hill. Available at <https://www.presidentialcabinets.org>.
- Comparative Agendas Project (2022). Comparative Agendas Project (CAP). Comparing Policies Worldwide. Research project, Comparative Agendas Project. Available at <https://www.comparativeagendas.net>.
- Eshima, S., Sasaki, T., and Imai, K. (2022). Keyword Assisted Topic Models. Package, Harvard University and Massachusetts Institute of Technology. Package 'keyATM' version 0.4.1.
- Greene, K. T., Park, B., and Colaresi, M. (2019). Machine Learning Human Rights and Wrongs: How the Successes and Failures of Supervised Learning Algorithms Can Inform the Debate About Information Effects. *Political Analysis*, 27(2):223–230. <https://doi.org/10.1017/pan.2018.11>.
- Lee, G. T., Kim, C. O., and Song, M. (2021). Semisupervised sentiment analysis method for online text reviews. *Journal of Information Science*, 47(3):387–403. <https://doi.org/10.1177/0165551520910032>.

References II

- Ooms, J. (2021). Open Source OCR Engine. Package, Berkeley Institute for Data Science. Package 'image.binarization' version 0.1.2.
- Sauvola, J. and Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236. [https://doi.org/10.1016/S0031-3203\(99\)00055-2](https://doi.org/10.1016/S0031-3203(99)00055-2).
- Singh, T. R., Roy, S., Singh, O. I., Sinam, T., and Singh, K. M. (2011). A New Local Adaptive Thresholding Technique in Binarization. *IJCSI International Journal of Computer Science Issues*, 8(6–2):271–277. <https://doi.org/10.48550/arXiv.1201.5227>.
- Su, B., Lu, S., and Tan, C. L. (2010). Binarization of historical document images using the local maximum and minimum. In *DAS '10: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 159–166, Boston. ACM Press. <https://doi.org/10.1145/1815330.1815351>.
- Watanabe, K. (2018). Newsmap. A semi-supervised approach to geographical news classification. *Digital Journalism*, 6(3):294–309. <https://doi.org/10.1080/21670811.2017.1293487>.
- Watanabe, K. and Zhou, Y. (2022). Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches. *Social Science Computer Review*, 40(2):346–366. <https://doi.org/10.1177/0894439320907027>.

Thank you very much!



Presentation compiled with **LATEX** and some 
 Download the latest version from [GitHub](#)